DOCUMENT RESUME

ED 123 260 TM 005 320

AUTHOR Morse, David T.; Morse, Linda W.
TITLE A Model for Assessing the Effects of Departures from Reality in Performance Testing.
PUB DATE [Apr 76]
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Cost Effectiveness; Decision Making; *Mathematical Models; Measurement Techniques; *Performance Tests; Statistical Analysis; *Test Construction; Testing Problems; *Test Reliability; *Test Validity; True Scores
IDENTIFIERS *Generalizability Theory

ABSTRACT

Performance testing often entails the usage of expensive, time-consuming measures in the quest for determining the level of performance on some desired behavior. It is concluded that a generalizability theory approach to dealing with departures from reality in testing can aid in the establishment of empirically-based choices of measurement strategies. This paper presents a model for assessing the loss of information due to using a measure which may be less realistic, but more feasible, than the desired behavior. The method is based on the concept of generalizability theory. An example is included along with a brief discussion of relevant considerations in performance testing, a background on generalizability theory, and a discussion on decision-making. (Author/DEP)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED123260

A Model for Assessing the Effects of Departures from Reality in Performance Testing

David T. Morse and Linda W. Morse

Career Education Center
Florida State University

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
DAVID T. MORSE
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

© David T. Morse and Linda W. Morse, April 1976.

TM005 320

A paper presented at the annual meeting of the American Educational Research Association, San Francisco, California, April 19-22, 1976.

A Model for Assessing the Effects of Departures from Reality in Performance Testing

David T. Morse and Linda W. Morse

Career Education Center
Florida State University

ABSTRACT

Performance testing often entails the usage of expensive, time-consuming measures in the quest for determining the level of performance on some desired behavior. This paper presents a model for assessing the loss of information due to using a measure which may be less realistic, but more feasible, than the desired behavior. The method is based on the concept of generalizability theory. An example is included along with a brief discussion of relevant considerations in performance testing, a background on generalizability theory, and a discussion on decision-making.

A Model for Assessing the Effects of Departures from Reality in Performance Testing[1]

The ideal measurement strategy in a performance-based learning situation is to have the learner attempt the desired behavior by demonstrating his competence in a performance setting. For instance, if the behavior is to "successfully overhaul and rebuild a V-8 engine," or to "successfully navigate on land from an unfamiliar point to base using only compass and relief map," the most desirable performance test for the first would be to supply an automobile with an engine in need of overhaul and supply the required tools and equipment. For the second example, the potential navigator should be placed in unfamiliar surroundings with only compass and relief map. Certain overriding considerations, however, may prohibit use of such direct measures of performance. For both examples, factors such as lack of time, money, equipment, and supervisory personnel might dictate that the test actually used be a measure as indirect as a short paper-and-pencil test covering selected aspects of engine rebuilding or land navigation. Clearly, this is not as desirable as use of the direct measure of performance. When decision-makers are confronted with the necessary use of a less direct measure of performance, however, how should they select which one to use, and how much loss of fidelity to the actual behavior must they accept? These questions should be asked and answered in situations where a less direct measure of performance is being utilized.

The purpose of this paper is to present a method of assessing the effects of departures from reality using a generalizability theory approach. Once the effects of changes in fidelity have been de-

termined, instructional designers and/or measurement specialists have a basis for the rational selection of measurement strategies. Further, if costs of testing and costs attached to losses of information concurrent with departures from reality can be determined, the selection of a measurement strategy can be based on a cost-effectiveness decision.

The remainder of this paper is divided into three sections: a brief discussion on contraints and fidelity in performance testing; a brief background on the rationale and mechanics of generalizability theory; and the methodology for the model, along with an example and discussion on decision-making.

Preliminary Considerations

Constraints in Performance Testing

A method for determining losses due to departures from reality in performance testing has not been adequately explored. This is an extremely crucial issue since the basic premise of performance testing lies in the measurement of presumably actual behaviors. Therefore, the only perfectly valid performance test would be one involving observation of the student's natural behavior. This would prove impossible in all but a few situations. Lindquist (1951) has identified several difficulties in direct measurement. First, the nature of the objective which is being assessed often makes it impossible to measure. Many objectives in the affective domain are examples of this situation. Secondly, a natural series of events may not be easily observable or either maybe inaccessible. Third, observing some behaviors may be exceedingly difficult or impossible because of the relative infrequency of occasions when the behavior

is naturally elicited. Lindquist also points out the problem of lack of comparability in accessible behavior samples for different students. Perhaps one of the most constraining obstacles in making direct measures lies in the difficulty of constructing such measures. Even simple performance objectives may yield complex behaviors which must be analyzed in order to develop appropriate and valid performance tests. The additional effort required for designing performance tests means their development is more costly in time and energy, yet they still may be plagued by one or more of these problems.

Lindquist outlines four basic types of tests: (a) giving the learner the opportunity on special occasion to perform the behavior specified in the objective; (b) having the student exhibit behavior(s) similar to the specified performance, making the assumption that a relationship exists between the behaviors desired and elicited; (c) giving the student a situation in which the desired behavior would be necessary and asking what should be done and/or how he would do it; and (d) testing the student on his or her knowledge of facts, rules, principles, etc. which are necessary for successful demonstration of the desired performance. These test types parallel succeeding levels of reality. For this paper, these four test types will be considered as: (a) actual; (b) simulated; (c) verbal; and (d) subordinate knowledge or skills. These last three types of measures represent departures from reality.

Although for many situations it is difficult to attempt to elicit the actual behavior, sometimes it is possible to do so. In Lindquist's

identical elements test, the elements of the actual performance must be identical to the critical elements in the criterion behavior even though they may be differently distributed in the natural or criterion situations. An example of this would be the applicant for a clerk/typist position who is asked to type a business letter as part of the job application process. This letter may differ in degree of difficulty from those typically typed on the job but the critical aspect of typing a letter in a business format is identical.

The second kind of test is the simulation. In a simulated test, the elements should be substantially related to the actual desired behavior. There should be considerable relationship between the elements in the simulation and in the actual test. This kind of test could be illustrated with the example of pilot simulator training machines.

The verbal description test type requires the student to respond to a situation based on how he would or ought to behave. The presentation of the situation may be oral or written and the pattern for response may vary from free response to selection between alternative answers. An example of this would be the vocational student who is presented with a situation describing a stalled car and is asked how he would diagnose the problem.

The fourth test type requires the student to exhibit his competence in a particular subject by demonstrating mastery of pertinent facts, rules, principles, etc. Although least desirable of the four types of tests in terms of fidelity, this format has been the most widely adapted type for measuring educational achievement. However, the demonstration of prequisite knowledge of a behavior

is not a sufficient condition for exhibiting a desired behavior due to large disparity between the two conditions.

Fidelity

Fidelity must be considered as the test designer moves from the real world to a simulated test environment. Fidelity is defined in terms of the degree of relationship between the real situation and the test conditions. This relationship is not entirely dependent on the face validity of the test conditions but instead depends on how well the skills and knowledge exhibited in the simulated (and lower) testing conditions transfer to the real world behavior (Branson, Rayner, & Epstein, 1974). However, one expects high fidelity when the test situation incorporates the highest level of reality possible (i.e., actual, simulated) and low fidelity with lower levels of measurement reality.

A valid performance test has been assumed to be one which has complete fidelity and comprehensiveness (Fitzpatrick & Morrison, 1971). But as tests more closely approximate the actual behavior they becomes harder to control because of the difficulty of observing the students under the same conditions. This difficulty over control leads to less reliable measures. Thus, it would appear that the more closely a test approximated the actual performance the difficulty with controlling the situation could cause a loss of reliability. This apparently paradoxical situation means that the dependability of different performance test scores from tests purporting to measure the same behavior may differ. Hence, the need for empirical determination of the interaction of degree of task fidelity and reliability of scores cannot be overemphasized if decisions are to be made from performance test results.

Rationale and Mechanics of Generalizability Theory

Classical Concepts of Reliability

The rationale underlying the concept of generalizability theory can be more easily understood in one is familiar with the classical notion of reliability. This definition of reliability of measurement is that of consistency or stability of a set of test scores. The important question in this definition lies in how the dimensions of score stability are interpreted in the traditional estimates of reliability. Before discussing the traditional reliability estimates, an overview of the types of variability which can affect test results should be outlined. Thorndike (1951) outlined the following:

1) Lasting and general characteristics of the individual
   (e.g., general level of intellect, ability to understand instructions)
2) Lasting but specific characteristics of the individual
   (e.g., knowledge of the subject specific to a set of test items)
3) Temporary but general characteristics of the individual
   (e.g., general state of health, fatigue, etc.)
4) Temporary but specific characteristics of the individual
   (e.g., subject interaction with a certain item or set of items)
5) Systematic or chance factors affecting the administration of the test or appraisal of test performance
   (e.g., noisy conditions for taking a test, a grader being given an incorrect answer key, etc.)
6) Chance or random variation
   (e.g., lucky guessing)

Depending on what is being measured, the sources of variation that should be accounted for should differ. Sources of variation included in the scores, but not measured as "true" variation introduce error into the measurement process.

Two traditional estimates of reliability, coefficient alpha and KR-20, are popular internal-consistency indices which tap the third source in Thorndike's list. That is, they reflect the degree to which a person's performance is consistent over a single set of items. Note that the variability attributable to sources 1 or 2 cannot be assessed using alpha or KR-20. Also, sources 1, 2, 4, 5, and 6 will be present in the set of scores, but alpha and KR-20 cannot detect them. Test-retest reliability considers the stability of scores across administrations of similar or alternate forms of a test. Thus, it is able to tap source 2. Sources 1, 2, 4, 5, and 6, however, will be present in the set of scores, but test-retest reliability will not be able to detect them. Reliability estimates derived from the Spearman-Brown prophecy formula can tap source 4. Although all the other sources of score variability may be present, the Spearman-Brown formula cannot detect them.

Thus, the particular estimate of reliability used can cause a difference in the estimated consistency of the scores. It is also likely that the characteristics of the examinees which should be measured are often not being measured as intended with these reliability estimates.

Rationale for Generalizability Theory

Instead of yielding a reliability coefficient, generalizability theory can yield a set of generalizability coefficients, and does so for an important reason. One is forced to question to what situation is he generalizing. That is, how consistent is a set of scores obtained under certain conditions? Here, the concept of universe and universe score is useful. Assuming that a population or domain of admissible observations of examinee performance can be defined, then this defined population constitutes the universe to which one could generalize. For instance, consider the following universe: Selected spelling words from the Kelly-James 10th grade spelling book, administered orally by one teacher on a Thursday afternoon in April. Assuming the measurements used were error-free, a true score could be obtained for each examinee for this universe. This score would be the examinee's universe score. Note that, for this example, if the universe of admissible observations is changed to include performance on two Thursdays in April, this would be a new universe, and each person could well have a different universe score. Thus, generalizability theory is concerned with the relationship of a set of observed scores to the corresponding universe scores for the examinees. The universe of conditions for performance assessment is necessarily specified. This is the fundamental difference between traditional reliability theory and the theory of generalizability. In generalizability theory, the universe that is being generalized to must be specified along with the admissible conditions of

observation for that universe. Hence, the sources of variation considered true variation, and the sources considered error are also specified.

Generalizability theory can help provide answers to a number of questions which a person using or building a test may ask. Some of the more fundamental of these questions are: (a) What is the examinee's universe score?; (b) What amount of error is there in the estimation of an examinee's universe score?; (c) What are the sources and relative sizes of variability in examinees' scores?; and (d) What changes can be made in the measurement process in order to reduce the error in estimating an examinee's universe score? Each of these questions will be discussed in greater depth later in this paper. The model in this paper draws from the work in generalizability theory by Cronbach et al. (1963; 1972).

Mechanics of Generalizability Theory

Conditions which serve to describe the universe of admissible observations are termed facets. Facets are analagous to factors in analysis-of-variance (ANOVA) designs. The basic determinations of generalizability analysis are achieved via an ANOVA approach. For example, consider a one-facet universe of different spelling words, with the population of words being all those in Webster's Third Edition. What this one-facet universe (i.e., of words, or items) means is that one is only interested in making an estimate as to how well a person can spell all the words in one dictionary, and this determination is made by observing performance over a single sample of

words from the dictionary. Suppose 100 words were randomly selected from the dictionary and administered to a group of 100 people. The resulting scores could be displayed in an array such as Table 1, below.

Table 1

Array of Hypothetical Administration of Spelling Words

		ITEMS							
		1	2	3	4	. . .	. . .	99	100
PERSONS	1	1	0	0	1			0	1
	2	0	1	0	0			0	1
	3	1	1	1	0			1	1
	. . .								
	98	0	0	0	0			1	0
	99	1	1	0	1			1	0
	100	0	1	0	1			0	1

Note: 1 indicates correct response, 0 indicates incorrect response

These results can then be analyzed in a two-way ANOVA design, in which persons and items could be set up as factors with no person-item replications. The ANOVA results would yield three distinct sources of variation: (a) variation attributable to persons; (b) variation attributable to items; and (c) residual variation. A sample table of output from an ANOVA analysis for this example is displayed in Table 2.

Table 2

Sample Output for Hypothetical Example

Source	SS	df	MS	E(MS)
Persons	193.05	99	1.95	$\sigma^2(res) + 100\sigma^2(P)$
Items	148.50	99	1.50	$\sigma^2(res) + 100\sigma^2(I)$
Residual	490.05	9801	.05	$\sigma^2(res)$
Total	831.60	9,999		

The coefficient of generalizability, or the relation between observed and universe scores is an intraclass correlation estimated by: $\hat{G} = \frac{\hat{\sigma}^2(P)}{\hat{\sigma}^2(P) + \hat{\sigma}^2(res)}$. The reader will note that these estimated variance components are derived using the E(MS)'s. In calculating $\hat{\sigma}^2(res)$, we can use MS(res) as an unbiased, maximum-likelihood estimate. Thus, $\hat{\sigma}^2(res) = MS(res) = 0.05$. The variance component for items is calculated in a similar manner: $\frac{MS(I) - MS(res)}{100} = \frac{1.50 - 0.05}{100} = \hat{\sigma}^2(I) = .0145$. Likewise, the variance component for persons is calculated as: $\frac{MS(P) - MS(res)}{100} = .0195$. The coefficient of generalizability is: $\hat{G} = \frac{.0195}{.0195 + 0.05} = \frac{.0195}{.0695} = .281$. This figure gives the estimated ratio of universe score variance to observed score variance, assuming the items selected are a random sample from the population of items, and the persons are representative of those which the scores are to be used for decision-making. For the one-facet case, using items as the facet, the coefficient of generalizability, here .28, is the same figure than would be obtained if the

data were analyzed for determining coefficient alpha or KR-20. Thus, alpha and KR-20 can be thought of as a special case of a generalizability coefficient for a universe of one facet. However, keep in mind the limitation that this coefficient refers only to administrations of similar sets of items to similar persons under exactly identical conditions. If the conditions of the test administration are to differ in the future, for instance, if tests are to be given before and after instruction, or at extremely different times of the day, or after long intervals of time, or given by a different teacher; any or all of these conditions might cause some variation in scores which will not be accounted for in the one-facet case. To remedy this, a multi-facet model is used. This is one advantage generalizability theory enjoys over classical reliability theory. To see the difference, we shall discuss a slightly more complex model.

Consider a two-facet model of alternate forms and occasions. That is, how stable are the spelling scores obtained if different sets of randomly selected words (the alternate forms) are used, and scores are taken across time (say, from week to week)? Once again, an ANOVA approach would be used. For simplicity, assume that each examinee takes each of three tests on each of three testing occasions one week apart, and order of tests is randomized. The analysis would proceed as though it would be done for a fully crossed factorial design. Total score variation in this example can be partitioned into seven components: (a) person variation; (b) test form variation; (c) occasion variation; (d) person X test interaction;

e) person X occasion interaction; f) test X occasion interaction; and g) residual variation. Thus, if three tests (10 items each) were administered on three different occasions to 100 examinees, the source table might look like the one presented in Table 3, below.

Table 3

Sample Output for Hypothetical Example

Source	SS	df	MS	E(MS)*
Persons	9875.55	99	99.75	$9\sigma^2(P) + 3\sigma^2(PT) + 3\sigma^2(PO) + \sigma^2(res)$
Tests	163.10	2	81.54	$300\sigma^2(T) + 3\sigma^2(PT) + 100\sigma^2(TO) + \sigma^2(res)$
Occasions	29.16	2	14.58	$300\sigma^2(O) + 3\sigma^2(PO) + 100\sigma^2(TO) + \sigma^2(res)$
P X T	1067.02	198	5.39	$3\sigma^2(PT) + \sigma^2(res)$
P X O	867.83	198	4.38	$3\sigma^2(PO) + \sigma^2(res)$
T X O	17.28	4	4.32	$100\sigma^2(TO) + \sigma^2(res)$
Residual	1537.67	503	3.06	$\sigma^2(res)$
Total	13,557.61	899		

*Using a random-effects model

The variance component estimates derived from this example are displayed in Table 4, below.

Table 4

Variance Component Estimates for Example

Component	Estimate of Variation	Proportion of total
$\sigma^2(res)$	3.06	.205
$\sigma^2(TO)$	0.013	.0008
$\sigma^2(PO)$	0.44	.029
$\sigma^2(PT)$	0.78	.052
$\sigma^2(O)$	0.03	.002
$\sigma^2(T)$	0.25	.017
$\sigma^2(P)$	10.34	.690

The estimate of the generalizability coefficient is:

$$\hat{G} = \frac{\hat{\sigma}^2(P)}{\hat{\sigma}^2(P) + \hat{\sigma}^2(PT) + \hat{\sigma}^2(PO) + \hat{\sigma}^2(res)} = \frac{10.34}{10.34 + .78 + .44 + 3.06} =$$

10.34/ 14.62 = .71. Thus, there is fairly good stability of scores across test forms and occasions. The relative sizes of the sources of variation (Table 4) show that tests, occasions, and their interactions account, in sum, for just over 10% of the total variation. The coefficient of generalizability would not be vastly different if the data were reanalyzed collapsing over: (a) occasions, making the one-facet model analagous to alternate forms reliability; or (b) tests, making the one-facet model analagous to test-retest reliability. Therefore, the two-facet model allows generalization to several universes--that of two facets, that of the first facet only, that of the second facet only, and variations on each, such as nested designs, fixed, and mixed models, and so on. The usage of generalizability theory allows much more flexibility in the analysis of performance assessments, thus more realistically reflecting the real world. There are many other considerations and analyses in generalizability theory, but for the purposes of this paper, we may stop at this point.

Description of the Model

Methodology and an Example

In evaluating alternative performance assessment strategies, a preliminary decision must be made concerning the face validity of each strategy. If a proposed alternative does not meet this first

requirement, then there is little value in attempting to use it. As an example, consider the skill of using an electric adding machine. While verbal aptitude test scores may correlate moderately with a person's facility in using an electric adding machine, few people would be satisfied with verbal aptitude tests as an alternative to performance assessment using an adding machine. Of course, the face validity determination should be made by those persons who have to make decisions about the examinees and/or learning situation.

After one or more alternate measures of the desired performance have been selected, carefully constructed, and tried out with a few representative examinees, a study of the alternate methods can be designed. If possible, the actual performance, or the simulation nearest to it should be included as one of the tasks. The purpose for inclusion of the actual performance is to provide scores which are as nearly error-free as possible, for determining the potential for misclassification in the alternate methods (this is discussed in more detail later), and for individual comparison of alternate strategies. In designing the study, the most powerful design is a fully crossed one, as in the examples discussed above. In the fully crossed design, each examinee is administered all tasks under all conditions deemed relevant enough to be included as a facet in the design. "Most powerful" refers to the precision of the variance component estimates, the G-coefficient estimate, and the error estimates. However, nearly any nested or mixed design can be utilized. If this is the case, however, some of the variance components which could be estimated in the crossed design may not be

directly estimable.

After the study is designed and carried out, the results should be analyzed and interpreted. Finally, the potential for misclassification should be considered under each alternative strategy. Costs for the alternate methods should also be taken into account. One method for using this information is presented by way of an example given below.

Consider the following example. The behavior of interest is to diagnose a fault requiring overhaul in a V-8 automotive engine and to overhaul and repair the engine. Relevant conditions might include: exercise to be completed within 150% of manufacturer's recommended flat-rate time; and each learner to execute the task alone. The criterion for successful performance is all operational checks on finished engine meeting manufacturer's specifications. Now, possible factors prohibiting such an exercise might be: lack of ample automobiles equally in need of engine overhaul; lack of up to forty hours "free time" for students to perform such an exercise; and lack of supervisory personnel to monitor many students. Hence, the case for some departure from reality is rather strong. Some reasonable alternative strategies might be: (a) allow the students to diagnose the fault in engines from five cars as well as describe the required repairs, or instead perform five small tasks involved in a complete engine overhaul on each of two engines; (b) verbally describe to an examiner the proper sequence of steps to follow when performing an engine overhaul; and (c) respond to a short-answer paper-and-pencil test composed of items dealing with diagnosis and

overhaul of an automotive engine. These alternatives correspond to levels b, c, and d of Lindquist's levels of measurement reality, respectively.

Suppose each of the ten learners selected were examined in each of the methods outlined above as well as being given a car in need of an engine overhaul and told to diagnose and repair the problem. Suppose further that the actual measure was scored 0 or 1, depending upon whether the rebuilt engine met all the manufacturer's operating specifications, alternatives a and b were scored 0 to 10, and the paper-and-pencil test was twenty items in length, each counting as one point. The order of these tasks could be randomly determined for the examinees so order effects would be minimized. Note that the assumption was made that all the alternatives met minimum face validity requirements. This describes a one-facet generalizability model. Suppose the score matrix in Table 5, below, resulted from the study.

Table 5

Hypothetical Score Matrix for Ten Examinees on Four Different Tasks*

Examinee/	Task 1	Task 2	Task 3	Task 4
1	1	10 (1)	10 (1)	20 (1)
2	0	5 (0)	5 (0)	15 (0)
3	1	9 (1)	10 (1)	18 (1)
4	0	2 (0)	4 (0)	12 (0)
5	0	1 (0)	2 (0)	17 (1)
6	1	10 (1)	9 (1)	16 (1)
7	1	10 (1)	10 (1)	20 (1)
8	1	10 (1)	10 (1)	20 (1)
9	1	4 (0)	5 (0)	16 (1)
10	0	8 (1)	8 (1)	16 (1)

*Numbers in parentheses are binary results given an arbitrary 80% criterion for "success" on each task. Task 1 is actual task, and tasks 2, 3, and 4 are alternative strategies a, b, and c, above.

Table 6, below, lists the results of a two-way ANOVA, using the binary scores from Table 5.

Table 6

Results of ANOVA

Source	SS	df	MS	E(MS)*
Persons	6.6	9	0.73	$\sigma^2(res) + 4\sigma^2(P)$
Tasks	0.3	3	0.10	$\sigma^2(res) + 10\sigma^2(T)$
Residual	2.2	27	0.08	$\sigma^2(res)$
Total	9.1	39		

*Using random-effects model

The resulting variance component estimates are listed below, in Table 7.

Table 7

Variance Component Estimates for Example

Component	Estimate	% of Total
Persons	.1625	66
Tasks	.002	01
Residual	.08	33

The G-coefficient, the measure of the degree of consistency of performance across tasks is: $\hat{G} = \hat{\sigma}^2(P)/(\hat{\sigma}^2(P) + \hat{\sigma}^2(res)) = .1625/.2425 = .67$. Since between-task variation accounts for only 1% of the total variation, if the cost of using the actual task is too great, the less realistic tasks could be used with little loss of information. The G-coefficient of .67 can be interpreted as the ratio of universe-

score variance to observed-score variance. Further analyses could be performed repeating the above, comparing individual tasks to the actual behavior. For instance, the G-coefficients for comparing alternate tasks two at a time are: .59, .59, and .82, corresponding to comparing tasks 1 and 2, 1 and 3, and 1 and 4, respectively. Interestingly enough, for this example, the usage of the paper-and-pencil test with an 80% criterion yields the least loss of information. The study could have been performed without setting criterion levels on the alternate tasks, and the resulting coefficients would not be drastically changed. The reason for the binary scores is for discussion of misclassification, in the next section. Finally, any study like this example would strive to include as many examinees as possible. The greater the number of examinees and levels of facets, the more dependable are the variance component estimates.

Errors and Decision-making with Results

The difference in an observed score X_{ij} for examinee i on task j and his universe score, μ_i (e.g., $X_{ij} - \mu_i$) is the error, Δ. Δ is analagous to the standard error of measurement in classical test theory. The size of the error Δ reflects the amount of information loss due to departures in task fidelity in the simulated tasks. A means for determining whether this loss is reasonable or not can be easily developed. First, the calculation of the error Δ should be explained. Since Δ reflects (average) within-person variation, Δ is calculated from the estimates of those variance components considered to be within persons. For the example, the calculation of Δ for the study is given in Table 8.

Table 8

Calculation of Error Δ

Variance component estimate	Observations within persons	Contribution to $\sigma^2(\Delta)$
$\hat{\sigma}^2(T) = .002$	4	.0005
$\hat{\sigma}^2(res) = .08$	4	.02
	$\hat{\sigma}^2(\Delta) =$	.0205
	$\hat{\sigma}(\Delta) =$	.145

Using this computation procedure, the expected size of $\sigma^2(\Delta)$ for any future study can be calculated. The same variance component estimates are used, and the (expected) number of within-person observations is used. For instance, for scores from five randomly-selected tasks, instead of four as in the example, the expected size of $\sigma(\Delta)$ is .13, about a 10% reduction. The same calculation procedure could be used with any number of different facets, although the variance-component estimates would be needed for the additional within-person facets.

The non-symmetrical nature of confidence intervals is aptly discussed by Cronbach et al. (1972), hence this discussion will only include the conservative Chebychev approach. For a randomly selected examinee in the example, a 75% confidence interval is given by $X_i \pm .3$, where $\sigma(\Delta)$ is used to obtain the .3.

Next, a threshold error level (TEL) has to be defined by the decision-makers. This is the size of the error Δ such that any values less than or equal to it are considered trivial, and any values greater are considered significant. The determination of the

TEL can be approached by setting a given level of savings desired in the total testing costs. That is, how much more economical does a procedure alternate to the actual performance test have to be in order to justify its use? (In a practical sense, this begs the question of not using the actual performance itself, for whatever the reason.) Suppose that in terms of personnel time alone, the cost was $20 per examinee to use the actual behavior for the performance test. For a group of ten examinees, the total cost is $200. Now, consider the cost of misclassification in terms of testing time alone. Both a false positive and false negative misclassification would require additional testing, but the false negative misclassification would constitute the only added cost, since the false positive misclassification would likely eventually have to be retested anyway. Suppose the cost for the most expensive alternate measurement strategy was $7.50 per man. Looking at the original score matrix (Table 5), a maximum of two misclassifications can be detected using any of the alternate tasks. Adding this to the original cost for testing ten persons makes the alternate task cost $90. Alternate task usage with the presently-set TEL (Δ) results in more than a 50% savings. As a general rule, therefore, once the desired amount of savings is determined (as long as it does not exceed an error-free cost of testing using the least expensive alternate measurement strategy), the TEL corresponding to that amount of savings can be compared with the observed size of the error Δ. If $\Delta \leq$ TEL, an alternate procedure is usable, and, according to this decision process, "reasonable." If $\Delta >$ TEL, then the number

of individual items required to reduce Δ to the TEL can be calculated, as explained above. This result is the desired length of the alternate task, or, in an analagous fashion, the number of alternate tasks which need be administered.

Summary

The advantage of the generalizability approach is obvious--not only can multiple levels of facets be considered simultaneously (something the product-moment correlation could not do), but it can incorporate multiple facets, and yields information on the relative sizes and sources of score variation. Also, between-person variation is not essential for useful results for decision-makers. The results of such an analysis can be used to aid in a rational, empirically-based decision for determining an appropriate measurement strategy. Cost-effectiveness decisions can also be made if the loss of information (expressed as the size of the error of measurement) due to different measurement approaches can be quantified on the same scale as the cost of testing.

For the field of performance testing, the authors conclude that a generalizability theory approach to dealing with departures from reality in testing can aid in the establishment of empirically-based choices of measurement strategies.

References

Branson, R. K., Rayner, G. T., & Epstein, K. The instructional systems development model: Design. (Phase II). Tallahassee, Fla.: Center for Educational Technology, Florida State University, Oct. 1974.

Cronbach, L. J, Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cronbach, L. J., Rajaratnam, W., & Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Fitzpatrick, R., & Morrison, E. J. Performance and product evaluation. In Thorndike, R. L., (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Lindquist, E. F. Preliminary considerations in objective test construction. In Lindquist, E. F., (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1971.

Thorndike, R. L. Reliability. In Lindquist, E. F., (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1971.

[1] Requests for copies of this paper may be sent to David T. Morse, Career Education Center, Florida State University, 415 N. Monroe Street, Tallahassee, Fla., 32306.